Міністерство освіти і науки України

Західноукраїнський національний університет

Факультет комп’ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

Звіт до лабораторної роботи №4

З дисципліни “Інтелектуальний аналіз даних”

Студента групи КН-31

Ковальковський В. В.

Тернопіль 2023

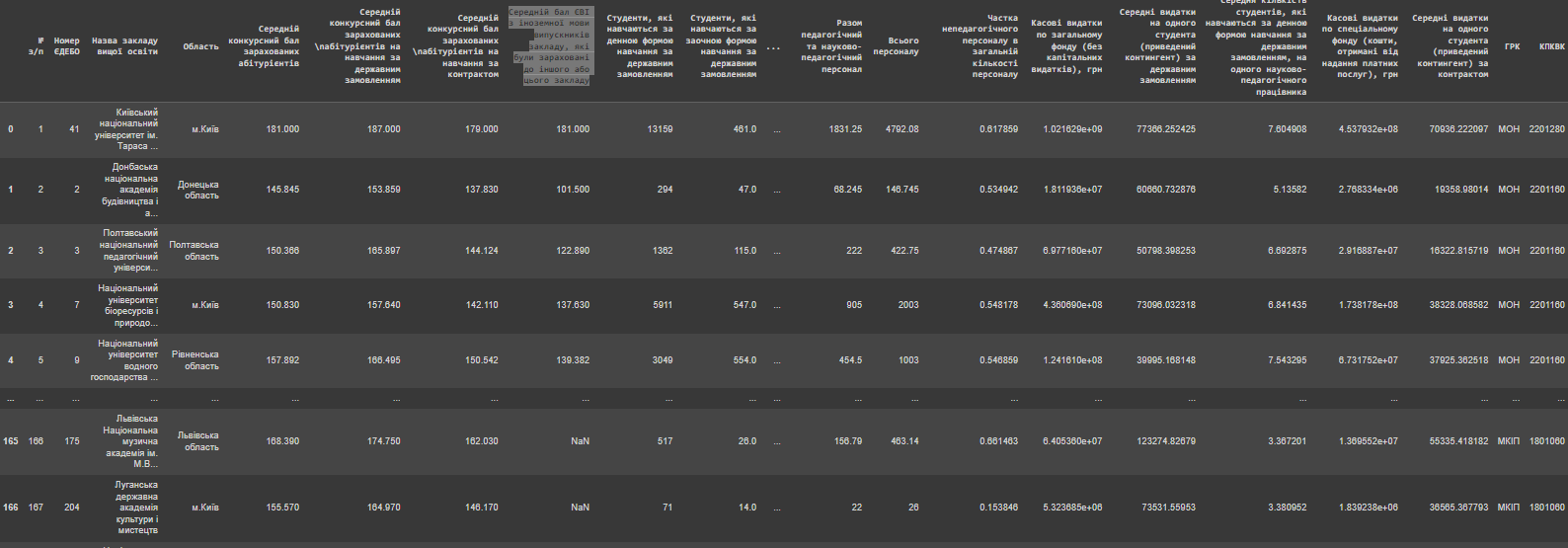
Тема. “ОСНОВИ СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ”

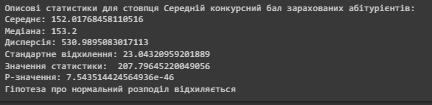
**Хід роботи**

**Варіант 14**

Витрати на ВНЗ за 2019

Загружаємо базу даних





За отриманими результатами можна зробити наступні висновки:

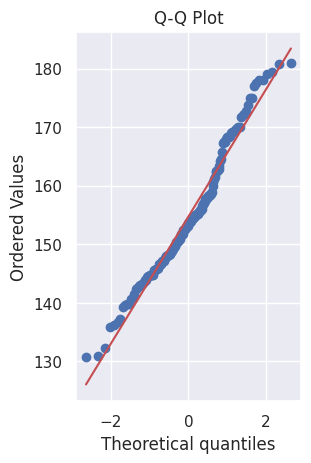
Середнє значення: 152.0177. Це середній конкурсний бал, який отримали абітурієнти, які були зараховані.

Медіана: 153.2. Медіана є значенням, яке розділяє розподіл пополам, і це показує, що половина абітурієнтів мали бали вище 153.2, а інша половина - нижче.

Дисперсія: 530.9895. Дисперсія вимірює розкид даних відносно середнього значення. Чим вище дисперсія, тим більший розкид в даних.

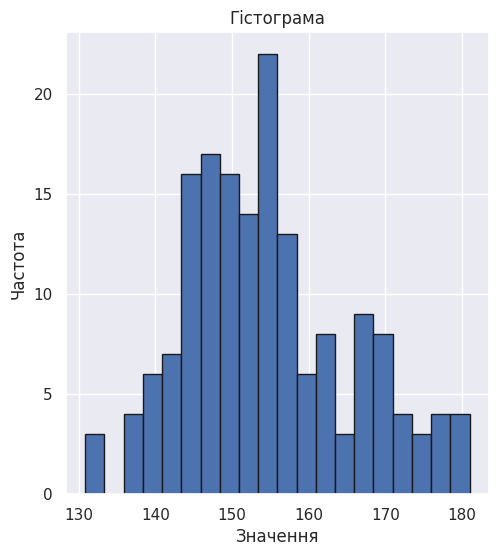
Стандартне відхилення: 23.0432. Стандартне відхилення є мірою розсіювання даних. В даному випадку, велике стандартне відхилення показує, що дані мають значний розкид навколо середнього.

P-значення: 7.5435e-46. P-значення вказує на ймовірність того, що спостережувані результати можуть виникнути випадково при певних умовах. У цьому випадку, P-значення дуже близьке до нуля, що свідчить про те, що гіпотеза про нормальний розподіл відхиляється на користь іншого розподілу даних.



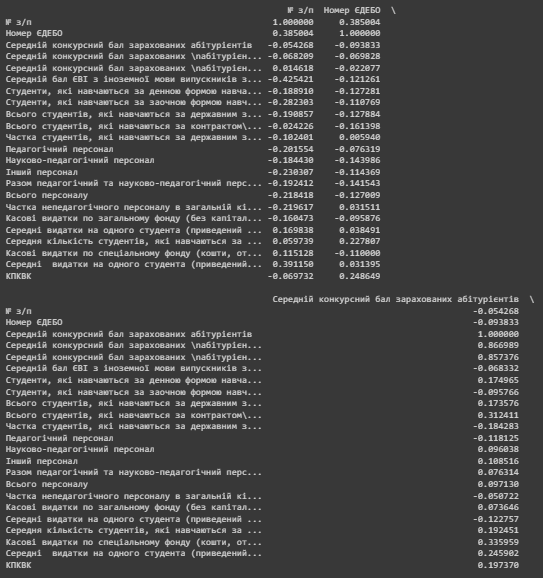
За наведеним графіком можна побачити, що точки лежать на лінії, яка представляє ідеальний нормальний розподіл. Це означає, що з урахуванням результатів графіку Q-Q і попередньої перевірки на нормальний розподіл,

можна зробити висновок, що дані мають сильно виражений нормальний розподіл.



Ця гістограма свідчить про те, що більшість найчастіше абітурієнти вступають з балами 145 – 157, високі бали (165-185) – низька тенденція, яка зустрічається набагато рідше.

Обчислюємо кореляцію різними змінними в наборі даних використовувати бібліотеку Python, таку як Pandas та NumPy. Існує кілька підходів до обчислення кореляції, таких як кореляція Пірсона, кореляція Спірмена та кореляція Кендалла



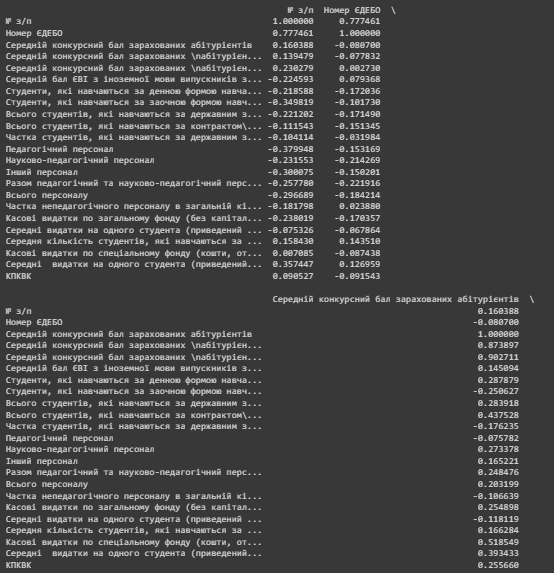
За наведеними вищще результатами кореляції Пірсона можна зробити висновок, що:

1. Значення кореляційного коефіцієнта між "№ з/п" і "Номер ЄДЕБО" дорівнює 0.385004. Це вказує на слабу позитивну кореляцію між цими двома змінними.

Значення кореляційного коефіцієнта між "Середній конкурсний бал зарахованих абітурієнтів" і "Середній бал ЄВІ з іноземної мови випускників з..." дорівнює -0.425421. Це вказує на помірну негативну кореляцію між цими двома змінними, що означає, що, якщо одна змінна зростає, інша зменшується, і навпаки.

Значення кореляційного коефіцієнта між "Середній бал ЄВІ з іноземної мови випускників з..." і "Студенти, які навчаються за заочною формою навч..." дорівнює -0.095766. Це вказує на дуже слабу негативну кореляцію між цими двома змінними, що може бути вважаною практично відсутньою.

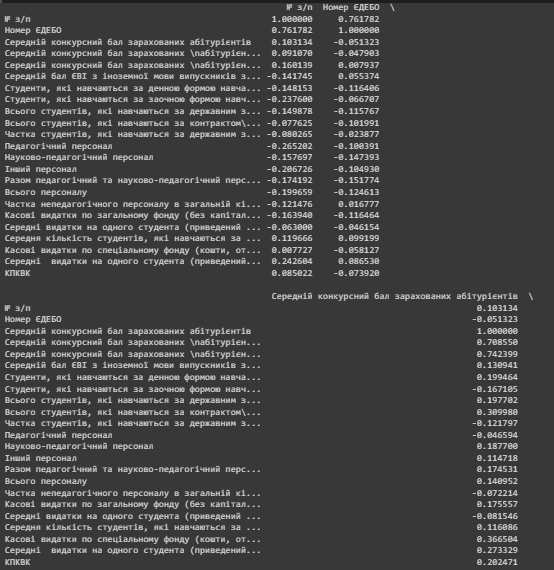
Значення кореляційного коефіцієнта між "Середні видатки на одного студента (приведений ..." і "Середня кількість студентів, які навчаються за ..." дорівнює 0.245902. Це вказує на помірну позитивну кореляцію між цими двома змінними.



За результатами кореляції Спірмена можна зробити наступні висновки:

1. Значення кореляційного коефіцієнта між "№ з/п" і "Номер ЄДЕБО" дорівнює 0.777461. Це вказує на дуже сильну позитивну кореляцію між цими двома змінними, що означає, що вони сильно зростають разом.
2. Значення кореляційного коефіцієнта між "Середній бал ЄВІ з іноземної мови випускників з..." і "Студенти, які навчаються за заочною формою навч..." дорівнює -0.349819. Це вказує на помірну негативну кореляцію між цими двома змінними, що означає, що, коли одна змінна зростає, інша зменшується.
3. Значення кореляційного коефіцієнта між "Середні видатки на одного студента (приведений ..." і "Середня кількість студентів, які навчаються за ..." дорівнює 0.166284. Це вказує на слабку позитивну кореляцію між цими двома змінними.
4. Значення кореляційного коефіцієнта між "Касові видатки по спеціальному фонду (кошти, от..." і "Середні видатки на одного студента (приведений..." дорівнює 0.393433. Це вказує на помірну позитивну кореляцію між цими двома змінними.

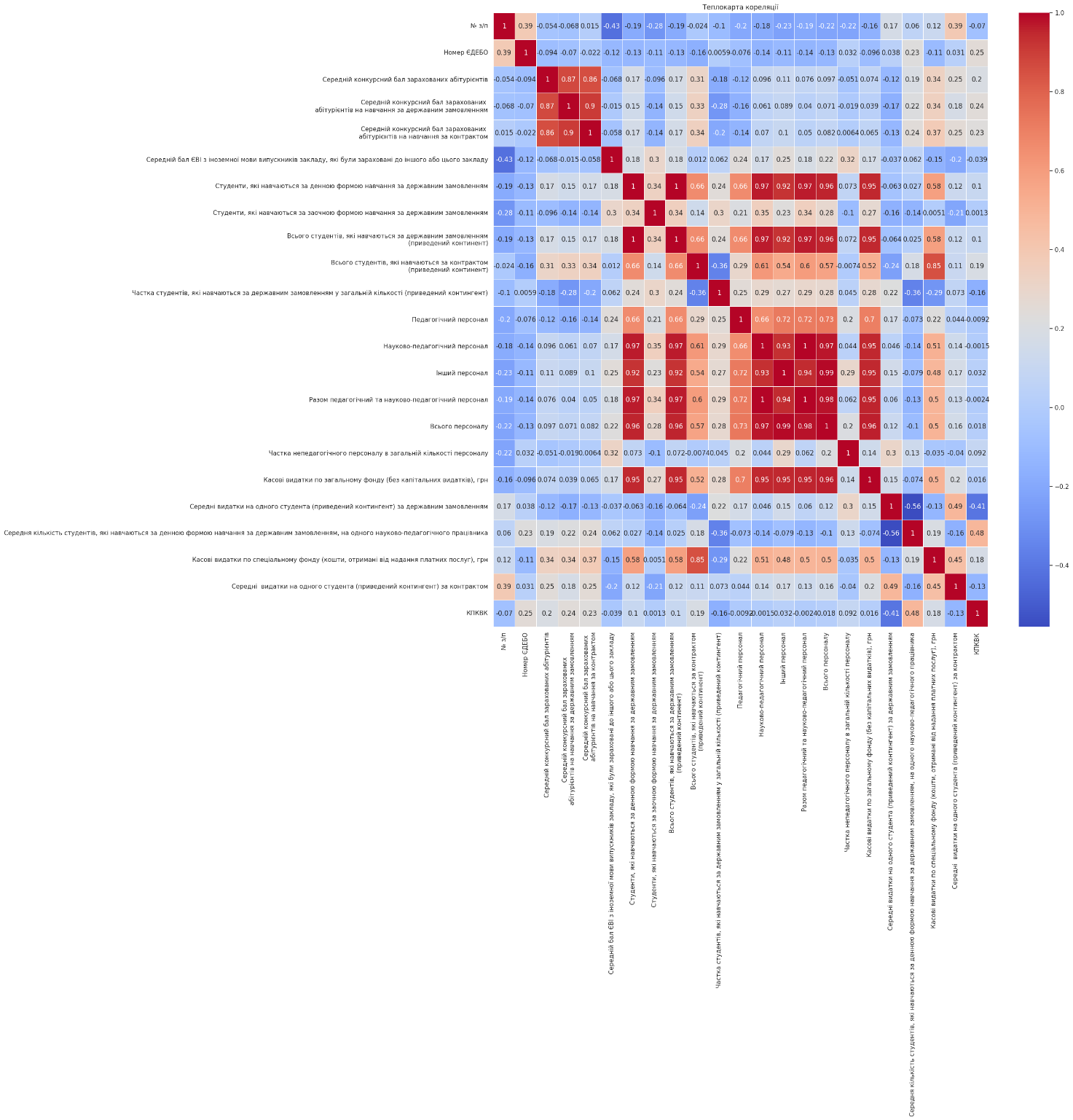
Кореляція Спірмена також може бути використана для визначення зв'язку між змінними, але вона вимірює не лінійний, а загальний зв'язок, що базується на рангах. У вашій таблиці видно сильні взаємозв'язки між деякими змінними, які можуть бути важливими для подальшого аналізу.



Основні висновки з результатів кореляції Кенделла в порівнянні з іншими методами:

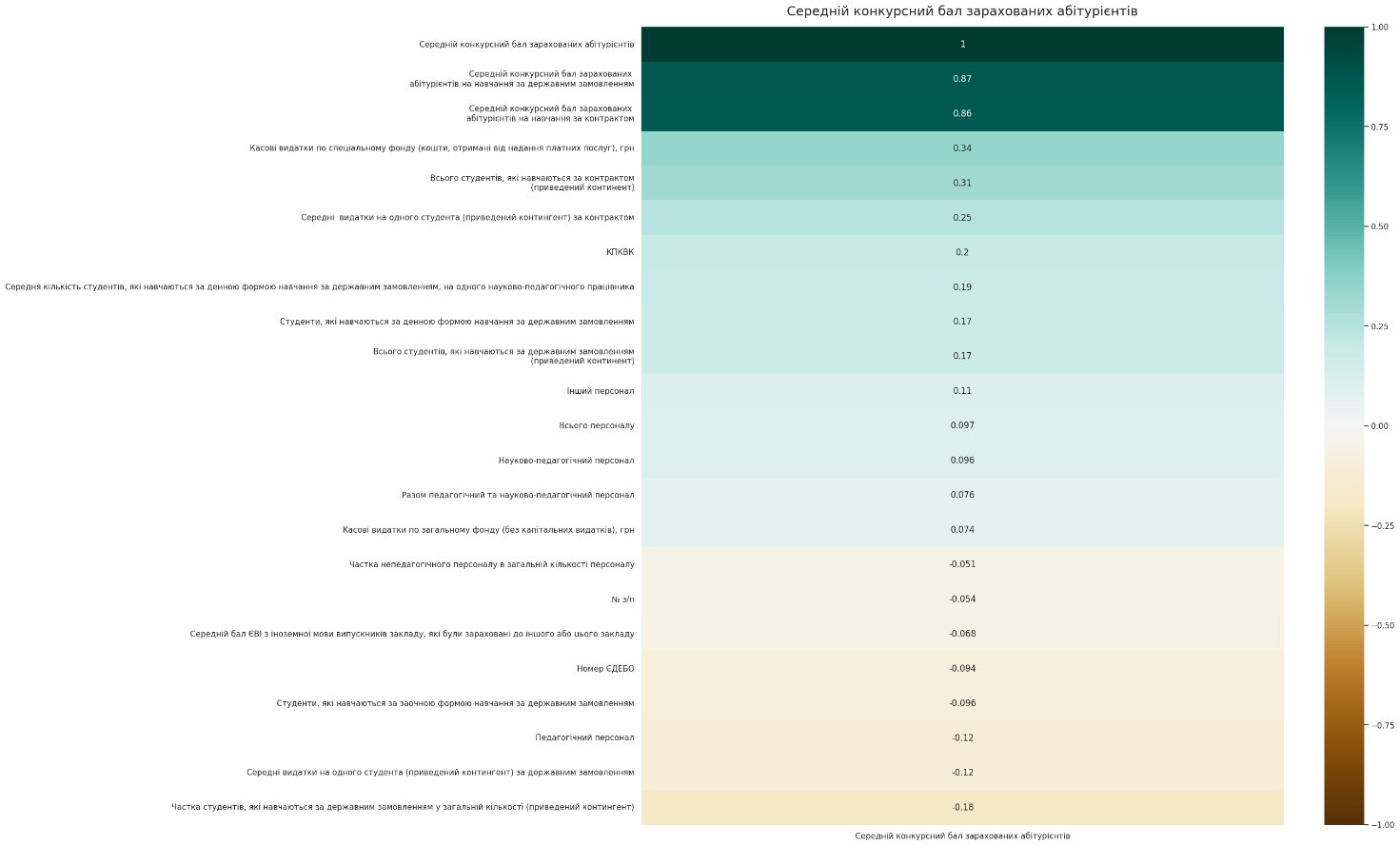
1. Значення кореляційного коефіцієнта Кендалла між "№ з/п" і "Номер ЄДЕБО" дорівнює 0.761782. Це вказує на високий ступінь конкордії між цими двома змінними, що означає, що, в основному, зростання в одній змінній супроводжується зростанням в іншій змінній.
2. Значення кореляційного коефіцієнта Кендалла між "Середній конкурсний бал зарахованих абітурієнтів" і "Середній бал ЄВІ з іноземної мови випускників з..." дорівнює -0.141745. Це вказує на слабку дискордію між цими двома змінними, що означає, що зростання в одній змінній супроводжується, в основному, спаданням в іншій змінній.
3. Значення кореляційного коефіцієнта Кендалла між "Касові видатки по спеціальному фонду (кошти, от..." і "Середні видатки на одного студента (приведений..." дорівнює 0.366504. Це вказує на помірну конкордію між цими двома змінними, що означає, що зростання в одній змінній супроводжується зростанням в іншій змінній.
4. Значення кореляційного коефіцієнта Кендалла між "Частка непедагогічного персоналу в загальній кі..." і "Касові видатки по спеціальному фонду (кошти, от..." дорівнює -0.072214. Це вказує на слабку дискордію між цими двома змінними, що означає, що зростання в одній змінній, в основному, супроводжується спаданням в іншій змінній.

Аналіз кореляції Кендалла допомагає зрозуміти, які змінні спільно змінюються та яким чином, враховуючи порядок даних, і часто використовується в статистичному аналізі для визначення взаємозв'язку між змінними без потреби в лінійних апроксимаціях.



У теплокарті кореляції кожен квадрат представляє кореляцію між двома змінними. Колір кожного квадрата відображає цю кореляцію: позитивну (більше 0) кореляцію може позначати яскравіший колір, негативну (менше 0) - темніший колір, а відсутність кореляції - блідий колір.

Зазвичай кореляція допомагає визначити, які змінні взаємозв'язані. Наприклад, якщо бачите сильну позитивну кореляцію між двома змінними, це означає, що зі зростанням однієї змінної інша теж зростає. На практиці, це означає, що зміни в одній змінній можуть впливати на іншу.

****

Створив теплову карту, яка візуалізує кореляцію між ознаками в DataFrame df та ознакою "Середній конкурсний бал зарахованих абітурієнтів" і виводить цю інформацію у вигляді кольорової картини з числовими значеннями кореляційних коефіцієнтів.

**Висновок:** під час виконання лабораторної роботи я освоїв навички опису даних проектування теплових карт, розрахунку кореляції та аналізу графіків.